MONITORING GEOGRAPHIC CONCENTRATION OF FEMALE BREAST CANCER USING CLUSTER ANALYSIS: THE CASE OF NEW MEXICO

F. Benjamin Zhan
Department of Geography
Southwest Texas State University
San Marcos, Texas 78666 USA
E-Mail: zhan@txstate.edu

1. INTRODUCTION

It is generally agreed upon that cancer is the combined result of personal genetic susceptibility and environmental conditions over time. Because of the role of environmental conditions in cancer, it is critically important that we detect possible geographic concentrations of a specific cancer as early as possible. One way to achieve this goal is to monitor cancer incidences periodically and detect possible statistically significant clusters of cancer in geographic space. In this discussion, the author presents a case study that demonstrates how to monitor geographic concentrations of female breast cancer in the state of New Mexico using spatial and space-time cluster analysis methods.

Female breast cancer incidence (morbidity) rates in Los Alamos County in New Mexico of the United States remained higher than the rest of the state for over two decades from 1970 to the early 1990s. There was an increase of breast cancer incidence rates between 1988 and 1992. (7) Past research concluded that these higher rates in Los Alamos County were attributed to reproductive and demographic factors, and the increase in breast cancer in the late 1980s and early 1990s was primarily due to increased detection of the disease at early stages. (7) However, because exposure to radiation is one of the primary causes of breast cancer (32), because of the close proximity of the communities with higher breast cancer rates to the Los Alamos National Laboratory, and because people living in New Mexico may have had exposure to atmospheric weapons testing, it is important to continue monitoring female breast cancer cases and detect possible geographic concentrations.

Past research compared breast cancer incidence rates in Los Alamos to those in reference populations selected in the state of New Mexico, and to those in the national statistics. (7) Although the merits of these studies cannot be denied, it remains unclear as to: (a) whether or not a statistically significant spatial cluster of breast cancer exists in populations living near Los Alamos, and (b) if a cluster exists, whether the cluster has persisted over time, particularly in recent years. Answers to these two questions are important because they will help address public concerns and develop possible preventive strategies of reducing breast cancer in communities near Los Alamos.

This study attempts to provide answers to the two questions mentioned above through both *spatial* and *space-time* cluster analyses using 17,119 individual breast cancer incident cases geographically-referenced to the county level in the entire state of New Mexico over a 25-year period from 1973 to 1997. The detection of cancer clusters in populations living near nuclear facilities has been an important research topic for

nearly two decades (20, 28), and many research activities have been carried out in countries like Canada (24), France (10, 30), Germany (12, 15), Spain (21), Sweden (11, 31), the United Kingdom (1, 3, 4, 5, 9, 27), and the United States (13, 22). Although results from most studies have been negative, evidence of excessive cancer rates have been found in places like Sellafield/Seascale (England) (4), Dounreay (Scotland) (5), La Hague (France) (30), and counties near Oak Ridge national Laboratory in Tennessee in the United States (22). But despite the fact that "strong links have been established between breast cancer risk and ionizing radiation [in women] (32)," very few research activities have been carried out to detect female breast cancer clusters in populations living near nuclear facilities. Monitoring geographic concentrations of female breast cancer in New Mexico will serve as an important step toward better understanding of the relationships between incidents of cancers and nuclear facilities in the United States.

2. DATA AND METHOD

A total of 17,119 female breast cancer incident cases in New Mexico over a 25-year period from 1973 to 1997 were extracted from the Public-Use CD-ROM (1973-1997) provided by the Surveillance, Epidemiology, and End Results (SEER) Program of the United States National Cancer Institute (NCI). Basic attributes associated with each record include a patient's county of residence at diagnosis and year of diagnosis. Covariates used in the analysis were race (white, black, and other) and age group (18 age groups with a five-year interval) at diagnosis. In addition to the incidence data, two additional data sets were also prepared. These two data sets were the at-risk population data and the geographic coordinates of county polygon centroids representing the locations of counties in New Mexico. The annual female at-risk population in each county in New Mexico over the 25-year period was obtained from the population database provided by the SEER Program of the NCI. To match the covariates in the incidence data, population counts for each combination of the three races (white, black, and other) and 18 age groups for each county in each year (54 combinations for each county in each year) were obtained.

Both spatial cluster and space-time cluster analyses of the incident cases were carried out using the SaTscan (Version 2.1.3) software package. This software package is based on the Spatial Scan and Space-Time Scan Statistic method developed by Kulldorff. (17) The software package was developed at and distributed by the NCI. A detailed description of the Spatial Scan Statistic is given in an article written by Kulldorff (17), and a description of the Space-Time Scan Statistic can be found in another article authored by Kulldorff and his colleagues. (19) The analytical procedure of the Spatial Scan Statistic consists of three steps. In the first step, the method draws circles centered at the centoid of each county in the study area in turn. At each county centroid, the sizes of the circles vary continuously from zero to a pre-specified upper limit (no larger than an area containing 50% of the at-risk population in the entire study area). The method then computes the numbers of cases inside and outside each circle and calculates the number of expected cases inside the circle based on the at-risk population in the area covered by the circle and the covariates used in the analysis. This process is repeated until all county centroids are scanned.

In the second step, the method determines the most likely cluster and secondary clusters. A cluster is composed of all adjacent counties the centroids of which fall within the same circle. The method achieves this objective by computing the likelihood ratio associated with a circle based on the parameter values obtained in previous steps. (17) If a circle has the maximum likelihood ratio and the number of cases within the circle is more than its corresponding expected number, then the cluster corresponding to this circle is considered to be the most likely cluster. Any other cluster with the number of cases within the cluster exceeding its corresponding expected number is considered a secondary cluster. Secondary clusters are less critical than the most likely cluster.

In the third step, the method evaluates the statistical significance of the most likely cluster and secondary clusters using Monte Carlo simulations. Under the null hypothesis, when cases are assumed to follow the Poisson distribution in space, no statistically significant spatial cluster exists. It follows that the simulated *p* value associated with the most likely cluster should be greater than a value indicating the significance level of the test (e.g., 0.05). Otherwise, the null hypothesis of randomly distributed cancer incident cases (no spatial cluster) is rejected and the most likely cluster is a statistically significant cluster. The significance of secondary clusters is evaluated in a similar manner.

Upon completion of the cluster analysis, the SaTscan software reports the most likely cluster, the secondary cluster(s), and counties within each cluster. In addition, the software reports the center and size of the circle corresponding to each cluster, the size of the at-risk population in a cluster, the number of observed and expected cases in a cluster, observed to expected ratio, log likelihood ratio, and *p* value associated with a cluster. For the space-time analysis, the time interval covered by a space-time cluster is also given.

The Spatial Scan Statistic method was used because it does not have the problem of multiple testing found in some exploratory analysis methods (8, 18, 25, 26), and because it does not require a user to specify the size and location of a cluster before the cluster analysis takes place. The Spatial Scan Statistic method evaluates all circles in the study area at once and uses the maximum likelihood ratio to select the most likely cluster and secondary clusters. It only evaluates statistical significance of the most likely cluster and secondary clusters, not clusters related to all circles. The exploratory analysis methods, however, evaluate the significance of the clusters related to all circles. Thus, they introduce the problem of multiple testing. The Spatial Scan Statistic method avoids the problem of multiple testing and is suitable for hypothesis testing. (18) The method proposed by Turnbull and his colleagues is similar to the Spatial Scan Statistic method, but it requires a user to specify the size of the cluster before the cluster analysis takes place. (18, 29)

Procedures associated with the Space-Time Scan Statistic are similar to those of the Spatial Scan Statistic except that circles in the Spatial Scan Statistic are replaced with cylinders. The circle at the base of a cylinder covers the spatial area of a potential cluster, and the height of a cylinder represents the time interval covered by the vertical dimension of the cylinder. Despite the fact that there is a number of methods for space-time cluster analysis (2, 6, 16, 14, 23), these methods do not possess the power to detect the size and location of a cluster simultaneously as well as evaluate the significance of a cluster. (18) In contrast, the Space-Time Scan Statistic can detect the location and size of the space-time clusters, and in the meantime, evaluate the statistical significance of the detected clusters. (18)

Both the spatial cluster analysis and the space-time cluster analysis were conducted on a Pentium III 866 MHz desktop computer. When running the program, the maximum size of a spatial cluster was set to contain 50% of the at-risk population and the maximum time interval in the space-time cluster was set to be 90% of the time span of the study period from 1973 to 1997. These maximum values ensure that the detected clusters, regardless of their location and size, are clusters detected without any pre-selection bias. It should be noted that the maximum allowed value of a spatial cluster does not mean that one has to pre-specify the size of a cluster before running an analysis. It simply means the largest allowed cluster would contain 50% of the at-risk population in the study area. This maximum value is reasonable because a cluster is expected to concentrate in certain area of the study region. If a cluster covers most of the study region, then the location and size of the study region is no longer meaningful in that study region. The Poisson probability model was used in the analyses. With 9,999 Monte Carlo simulations, it took the program 1 minute and 6 seconds to complete the spatial cluster analysis and 5 minutes and 1 second to complete the space-time cluster analysis.

3. RESULTS

The detected clusters and their associated parameter values are given in Table 1. Only one cluster was detected in the spatial cluster analysis and this cluster is statistically significant (p=0.0001) (Figure 1). This spatial cluster is centered in Santa Fe County and it contains four counties including Santa Fe, Los Alamos, Sandoval, and Bernalillo. A statistically significant space-time cluster (p=0.0001) with the same spatial coverage for the time period from 1984 to 1997 (inclusive) was detected in the space-time cluster analysis (Table 1; Figure 2). This space-time cluster is also centered in Santa Fe County. One secondary space-time cluster was detected, but this cluster is not statistically significant. This secondary space-time cluster contains Eddy County for the year 1995 and has a p value of 0.9873.

TABLE 1
CLUSTER ANALYSIS RESULTS OF FEMALE BREAST CANCER IN NEW MEXICO, 1973-1997. (Note: Incidence rates were adjusted for race and age group. Clusters with their associated p-values underlined are statistically significant clusters.)

Cluster category	Counties within the cluster	Time period	No. of observed cases	No. of expected cases	Observed to expected ratio	Log likelihood ratio	p Value
Spatial cluster analysis							
Most likely cluster	Los Alamos, Sandoval, Santa Fe, Bernalillo	NA	9,029	7,563.67	1.194	252.16	0.0001
Space-time cluster analysis							
Most likely cluster	Los Alamos, Sandoval, Santa Fe, Bernalillo	1984-97	6,730	4,979.95	1.351	409.42	0.0001
Secondary cluster	Eddy	1995	53	33.48	1.583	4.84	0.9873

The number of observed and expected incident cases in each county within the most likely spatial cluster over the 25-year period was also obtained (Figure 1). It is clear from these results that the number of observed cases in Los Alamos County is 48.7% more than the number of expected cases. The rate of excessiveness in other counties in the cluster is less severe, with 24.7% more than expected in Sandoval County, 18.7% in Bernalillo County, and 14.3% in Santa Fe County.

4. DISCUSSION

The results of this analysis give a clear picture of the geographic concentration of female breast cancer incident cases in New Mexico and provide answers to the two questions raised at the beginning of this discussion. For the first question, it is apparent from the results that a statistically significant spatial cluster of female breast cancer exists in populations living in the four-county area of Los Alamos, Sandoval, Santa Fe, and Bernalillo. For the second question, results from the space-time cluster analysis indicate that this spatial cluster is statistically significant for the period from 1984 to 1997, which means that the excessive breast cancer rates persisted until recent years in the four counties. Among the four counties in the most likely cluster, Los Alamos clearly has a higher degree of excessiveness of breast cancer with 48.7% more observed cases than expected. For the rest of New Mexico, no statistically significant cluster was detected. These results would help public health professionals form new hypotheses that can be used to study the causes behind the formation of the clusters.

FIGURE 1

COUNTIES IN THE SPATIAL CLUSTER OF FEMALE BREAST CANCER IN NEW MEXICO, 1973-97. (Note: incidence rates were adjusted for race and age group. The pair of numbers next to each county name is the observed/expected number of cases in that county.)

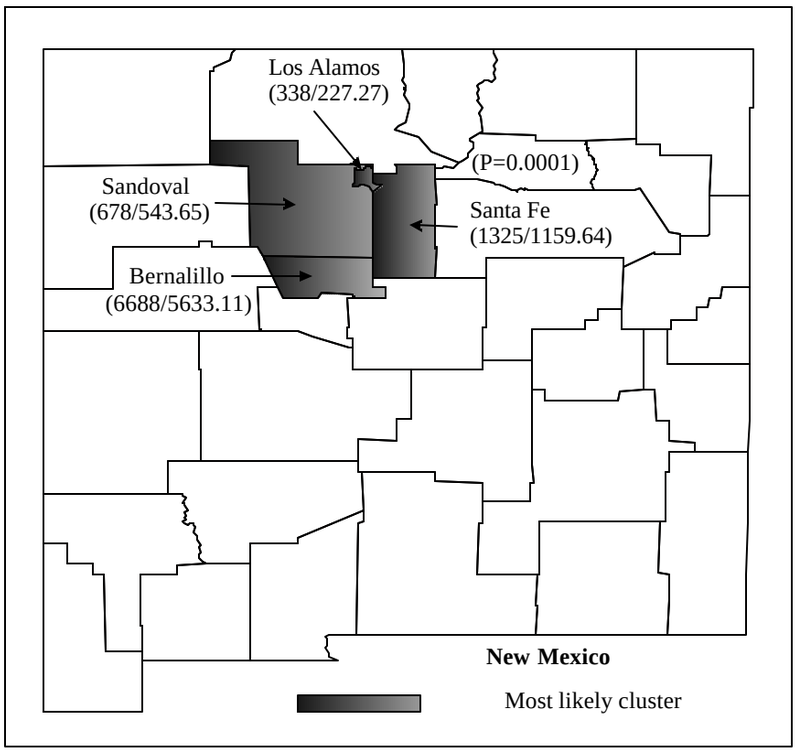
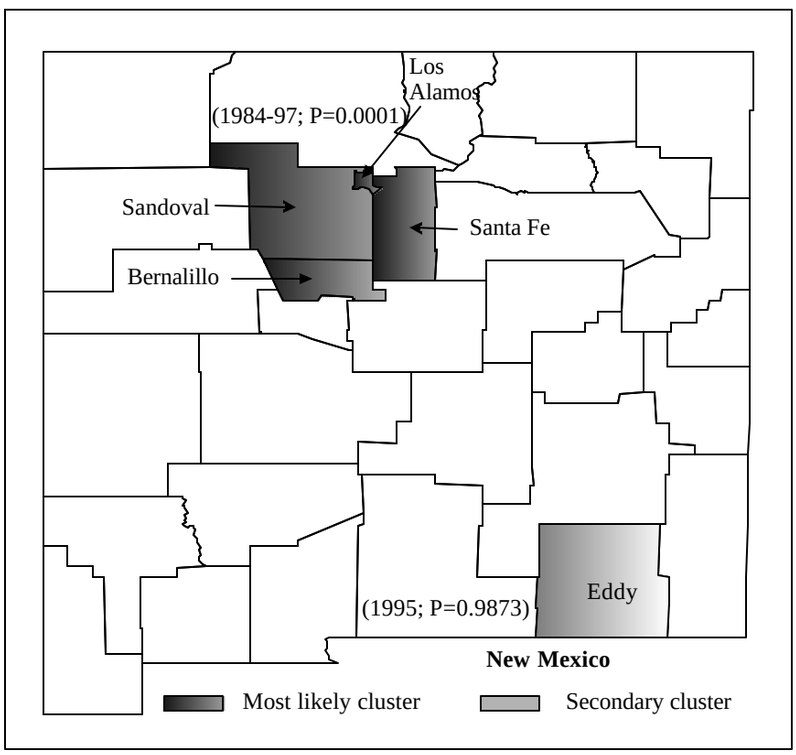


FIGURE 2

COUNTIES IN THE SPACE-TIME CLUSTERS OF FEMALE BREAST CANCER IN NEW MEXICO, 1973-97. (Note: incidence rates were adjusted for race and age group. The secondary cluster is not statistically significant.)



Future research needs to be directed at evaluating the reasons behind the formation of the female breast cancer cluster in populations living near Los Alamos. There are a few critical questions that remain to be answered. First, is there any possibility that the cluster is associated with the nuclear atmospheric testing research facilities in New Mexico? If so, what are the potential processes that have contributed to the formation of the female breast cancer cluster in New Mexico, and how would these processes function in geographic space and time and lead to the formation of the cluster?

Second, if the cluster is not associated with the nuclear atmospheric testing research facilities in New Mexico, what could be the other social and environmental processes or agents that may have contributed to the formation of the cluster, and how would these processes and agents function in geographic space and time? For instance, it is possible that different social and environmental processes and/or the interaction of some social and environmental processes may have contributed to the formation of the cluster of female breast cancer in populations living near Los Alamos over time. Finding answers to these questions will require long-term work, and it is a subject for future research.

5. ACKNOWLEDGMENT

Data used in the analyses reported in this article were from Surveillance, Epidemiology, and End Results (SEER) Program Public-Use CD-ROM (1973-1997), National Cancer Institute, DCCPS, Cancer Surveillance Research Program, Cancer Statistics Branch, released April 2000, based on the August 1999 submission. The author wishes to thank a reviewer for the helpful comments on an earlier version of this article.

6. REFERENCES

1. Alexander, F.E., R. Cartwright, P.A. McKinney, and T.J. Ricketts. 1990. Investigation of spatial clustering of rare diseases: Childhood malignancies in North Humberside. Journal of Epidemiology and Community Health 44(1):39-46.

2. Baker, R.D. 1996. Testing for space-time clusters of unknown size. Journal of Applied Statistics 23:543-554.

3. Bithell, J.F., S.J. Dutton, G.J. Draper, and N.M. Neary. 1994. Distribution of childhood leukemias and non-Hodgkin lymphomas near nuclear installations in England and Wales. British Medical Journal 309:501-511.

4. Black, D. 1984. *Investigation of the possible increased incidences of cancer in West Cumbria*. London, United Kingdom: Her Majesty's Stationary Office.

5. Black, R.J. and L. Sharp. 1993. Leukemia and non-Hodgkin's lymphoma: Incidence in children and young adults resident in the Dounreay area of Caithness, Scotland in 1968-1991. Journal of Epidemiology and Community Health 48(3):232-6.

6. Diggle, P, A.G. Chetwynd, R. Haggkvist, and S.E. Morris. 1995. Second-order analysis of space-time clustering. Statistical Methods in Medical Research 4:124-136.

7. DOE (U. S. Department of Energy). 1996. Final Programmatic Environmental Impact Statement for Stockpile Stewardship and Management, 0236 EIS, Volume II, Appendix E.4 Health Effects Studies: Epidemiology, http://search.dis.anl.gov/fpims_v2.0/eis/eis-0236/0236vol2/0236appe/e4.htm#e46. (Last test access on June 1, 2001)

8. Fotheringham, A.S., and F.B. Zhan. 1996. A comparison of three exploratory methods for cluster detection in spatial point patterns. Geographical Analysis 28(3):200-218.

9. Gardner, M.J. 1989. Review of reported increases of childhood cancer rates in the vicinity of nuclear installations in the UK. Journal of the Royal Statistical Society B 152(3):307-325.

10. Hattchouel, J.M., A. Laplanche, and C. Hill. 1995. Leukemia mortality around French nuclear sites. British Journal of Cancer 71(3):651-3.

11. Hjalmars, U., M. Kulldorff, G. Gustafsson, and N. Nagarwalla. 1996. Childhood leukemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. Statistics in Medicine 15(7-9):707-15.

12. Hoffman, W., H. Dieckmann, H. Dieckmann, and I. Schmits-Feuerhake. 1997. A cluster of childhood leukemia near a nuclear reactor in Northern Germany. Archives of Environmental Health 52:275-6.

13. Jablon, S., Z. Hrunec, and J.D. Boice. 1991. Cancer in populations living near nuclear facilities: A survey of mortality nationwide and case in two states. Journal of the American Medical Association 265(11):1403-8.

14. Jacquez, G.M. 1996. A k nearest neighbor test for space-time interaction. Statistics in Medicine **15**:1935-1949.

15. Kaatsch, P., U. Kaletsch, R. Meinert, and J. Michaelis. 1998. An extended study on childhood malignancies in the vicinity of German nuclear power plants. Cancer Causes and Control. 9(5):529-533.

16. Knox, G. 1964. The detection of space-time interactions. Applied Statistics 13:25-29.

17. Kulldorff, M. 1997. A spatial scan statistic. Communications in Statistics: Theory and Methods 26:1481-1496.

18. Kulldorff, M. 1998. Statistical methods for spatial epidemiology: Tests for randomness. In *GIS and Health*, ed. M. Loytonen and A. Gatrell, 49-62. London, England: Taylor & Francis.

19. Kulldorff, M., F.A. William, E.J. Feuer, B.A. Miller, and C.R. Key. 1998. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. American Journal of Public Health 8(9):1377-1380.

20. Laurier, D. and Bard. 1999. Epidemiologic studies of leukemia among persons under 25 years of age living near nuclear sites. Epidemiologic Reviews 21:188-206.

21. Lopez, A.G., N. Aragones, M. Pollan, M. Ruiz, and A. Gandarillas. 1999. Leukemia, lymphomas, and myeloma mortality in the vicinity of nuclear power plants and nuclear fuel facilities in Spain. Cancer Epidemiology, Biomarkers and Prevention 8(10):925-34.

22. Mangano, J.J. 1994. Cancer mortality near Oak Ridge, Tennessee. International Journal of Health Services 24(3):521-533.

23. Mantel, N. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. Cancer Research **27**:209-220.

24. McLaughlin, J.R., E.A. Clarke, E.D. Nishri, and T.W. Anderson. 1993. Childhood leukemia in the vicinity of Canadian nuclear facilities. Cancer Causes and Control 4(1):51-8.

25. Openshaw, S., M. Charlton, C. Wymer, and A.W. Craft. 1987. A mark 1 analysis machine for the automated analysis of point data sets. International Journal of Geographic Information Systems 1:335-358.

26. Rushton, G., and P. Lolonis. 1996. Exploratory spatial analysis of birth defect rates in an urban population. Statistics in Medicine 7:717-726.

27. Sharp, L., P.A. McKinney, and R.J. Black. 1999. Incidence of childhood brain and other non-haematopoietic neoplasms near nuclear sites in Scotland, 1975-94. Occupational and Environmental Medicine 56(5):308-14.

28. Shleien, B., A.J. Ruttenber, and M. Sage. 1991. Epidemiologic studies of cancer in populations near nuclear facilities. Health Physiology 61(6):699-713.

29. Turnbull, B., E.J. Iwano, W.S. Burnett, H.L. Howe, and L.C. Clark. 1990. Monitoring for clusters of disease: Application to leukemia case in upstate New York. American Journal of Epidemiology 132:S136-S143.

30. Viel, J.F., D. Pobel, and A. Carre. 1995. Incidence of leukemia in young people around the La Hague nuclear waste reprocessing plant: A sensitivity analysis. Statistics in Medicine 14(21-22):2459-72.

31. Waller, L.A., B.W. Turnball, G. Gustafsson, U. Hjalmars, and B. Anderson. 1995. Detection and assessment of clusters of disease: An application to nuclear power plant facilities and childhood leukemia in Sweden. Statistics in Medicine 14(1):3-16.

32. Wolff, M,S,, G.W. Collman, J.C. Barrett, and J. Huff. 1996. Breast cancer and environmental risk factors: epidemiological and experimental findings. Annual Review of Pharmacology and Toxicology 36:573-96.